Big data Analysis

Real time data processing

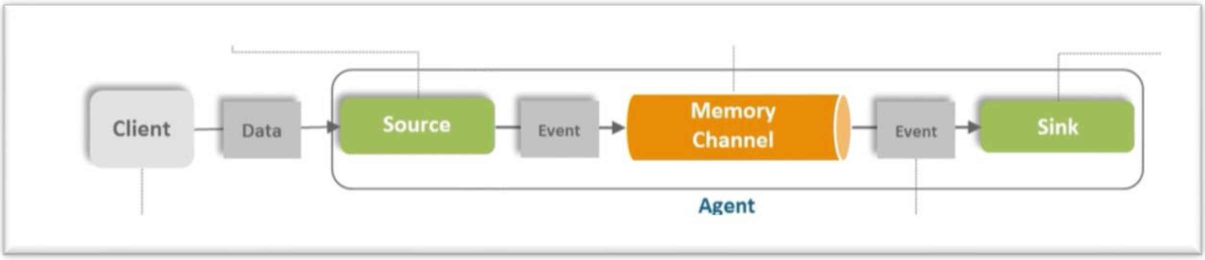
Module – 8

Homework

1. **What is Flume?**

   Flume is the distributed, reliable and available service for efficiently collecting, aggregating, and moving large amounts of log data from various sources to a centralized data store.

2. **Explain the core components of Flume.**



3. **What is an Agent?**

   Any physical java virtual machine that running the flume is called agent. It is a collection of sources, sink and channel.

4. **What is a channel?**

   The conduit between the source and the sink is called channel. Sources ingest events into the channel and the sinks drain the channel.

5. **What is Kafka?**

   Apache Kafka is a distributed publish-subscribe messaging system. It was originally developed at LinkedIn and later on became a part of Apache project. Kafka is fast, scalable, durable, fault-tolerant and distributed by design.

6. List the various components in Kafka.

The main Kafka components are topics, producers, consumers, consumer groups, clusters, brokers, partitions, replicas, leaders, and followers.

7. What is the role of the Zookeeper?

Zookeeper service is mainly used for coordinating between brokers in the Kafka cluster. Kafka cluster is connected to Zookeeper to get information about any failure nodes.

8. Why are Replications critical in Kafka?

The purpose of adding replication in Kafka is for stronger durability and higher availability. We want to guarantee that any successfully published message will not be lost and can be consumed, even when there are server failures.